Automated Glaucoma Detection using Structural Optical Coherence Tomography with Data Mining

[1]Saksham Sood, [2]Aparajita Shukla, [3]Arif Arman, [4]Poonkodi.M

[1,2,3]Students, Department of Computer Science, SRM University, Chennai, INDIA

[4]Assistant Professor, Department of Computer Science, SRM University, Chennai, INDIA

***Abstract*— Glaucoma is an estate that causes damage to your eye's optic nerve and gets worse over time. Detection of glaucoma at the earlier stages is almost impossible and hence automation detection of glaucoma is very important and needs to be improved. However, there are many detection techniques which are used for latter stage, one such technique is Visual Field test (Perimetry test) which helps to find certain patterns of vision loss. This may mean a certain type of eye disease is present. In the Visual Field Perimetry test a detailed record of your visual fields are made which includes baseline information, descriptions, or drawings. This information which is collected can be compared with a normative database containing the measurements of an unaffected person. This paper proposes a method through which one can predict the Perimetry Test results using OCT (Optical Coherence Tomography) which can be assumed as a more reliable technique because of its structural analysis and lack of human error(s). OCR is used for extracting the data from OCT test results(including OCT Macular Thickness Map measurements) and then Data pre-processing is done on raw data to prepare it for another processing procedure followed by regression technique of data mining to explore the data in search of consistent patterns and systematic relationships. Lastly, Cross validation is done for the assessment of results to obtain statistical analysis.**

Keywords—Data Mining; Glaucoma Detection; Glaucoma Prediction; Feature Selection; OCR

I. INTRODUCTION

Glaucoma is a disease of optic nerve damage and is a leading cause of vision loss. Currently, Glaucoma is detected by using VF (Visual Field) test usually which may include human errors and so is not said to be reliable. Therefore, a more reliable approach is used. Optical coherence tomography (OCT) [5] is a medical imaging technique (OCT macular thickness measurements) which provides an important structural analysis in current clinical management system.

Optical character recognition (OCR) is the translation of scanned documents into machine-encoded text. The OCR conversion process needs to have an OCR Engine configured along with it to run successfully. The OCR Engine can extract the text from the screenshot of a control. One of the most popular and accurate yet Open Source OCR Engine is Tesseract which is used in the implementation part of this paper. Data Mining is a process of extracting some trivial information/relationship from a large static database.

It is the analysis step of the Knowledge Discovery Database which is very useful in directing the most important functionalities hidden in a large data warehouse. 100 cases were used (combining both patients diagnosed with (and) without glaucoma) in order to discover the trivial information by applying data mining.

Compared with the usual methods of testing, a mathematical model based on the historical data, cross-validation is a statistical procedure which produces an estimate of forecast skill which is less biased. It is a technique which enhances the existing best models using all of the data and avoids the issue of over-fitting. In addition it helps to assess the predictive accuracy of a model in a sample taken for the test. A commonly used technique called the 10-fold cross-validation technique is used in this paper which is best suited for this purpose because it has low variance and limited data.

Many efforts have been put forth in the detection of glaucoma and few are explained below-

1. AUTOMATED GLAUCOMA DETECTION PROCESS

Automated glaucoma detection model [1] consists of step by step process that starts with the collection of retinal images using digital image capturing devices like Pictor, Retcam Shuttle etc. followed by Pre-processing so as to equalize the irregularities existing in the images. Then feature extraction is done simplifying the amount of resources required to describe a large data set accurately. There are various feature extraction techniques such as Moment Method, FMM in painting Method, Macular cube algorithm, Histogram Model, P-tile threshold Method, FFT coefficients etc. Lastly, Classification is performed in which a significant set of data that can be further classified is analyzed and categorized as Normal or Glaucoma effected.

2. CONFOCAL SCANNING LASER OPTHALMOLOGY (CSLO)

Confocal Scanning Laser Opthalmology [6] is an image gaining technology which is proposed to improve the quality of the examination compared to ordinary ophthalmologic examination using laser. Laser scans the retina along with a detector system and once a single spot on the retina is irradiated, a high-contrast image of great reproducibility can then be used to estimate the width of the RETINAL NERVE FIBRE LAYER (RNFL).

The Heidelberg Retinal Tomography is possibly the most common example of this technology.

3. SCANNING LASER POLARIMETRY (SLP)

This [8] method uses a 780-nm diode to illuminate optic nerve. The evaluation of the emerging polarized state of light is linked with RNFL thickness. RNFL causes a change in the state of divergence of a laser beam as it passes. Unlike CSLO, SLP can directly measure the thickness of the RNFL. GD is an example of a scanning laser polarimeter. It contains a normative database and statistical software package to permit comparison to age-matched normal subjects of the same racial origin. The advantages of this system are that images can be obtained without pupil dilation, and evaluation can be done roughly in 10 minutes. Modern instruments have added improved and erratic corneal compensation technology to account for corneal polarization

4. WAVELET FOURIER ANALYSIS (WFA)

It is a Texture-based technique [2] which has been proven successful but is still a challenge to generate features that retrieve generalized structural and textural features from retinal images. To overcome the generalization of features wavelet transforms in image processing are used to extract the texture features from images. In wavelet transform, the image is represented in terms of the frequency of content of local regions over a range of scales. This representation helps the analysis of image features, which are independent in size and can often be characterized by their frequency domain properties. The use of WFA for the categorization of neuroanatomic distraction in glaucoma has achieved substantial success. WFA is used as an exact model to analyze and parameterize the temporal, superior nasal, inferior, and temporal shapes. Two types of wavelet transforms are used, discrete wavelet transforms and continuous wavelet transforms in image processing. In this method, discrete wavelet transform using a fourth-order symlets wavelet is used to extract features and analyze discontinuities and rapid changes contained in signals. Discrete wavelet transform is a multiscale analysis method where analysis can be performed on a range of scales. Each level of the transformation provides an analysis of the source image at a different resolution, resulting in its independent rough calculation and detailed coefficients. In the WFA, the fast Fourier transform is applied to the detailed coefficients. The resultant Fourier amplitudes are combined with the normalized approximation coefficients of the DWT to create a set of features. Quantitatively examine the effectiveness of different wavelet filters on a set of curated glaucomatous images by employing the standard 2-D-DWT. Use three well-known wavelet filters, the daubechies (db3), the symlets (sym3), and the biorthogonal (bio3.3, bio3.5, and bio3.7) filters. Then calculate the averages of the detailed horizontal and vertical coefficients and wavelet energy signature from the detailed vertical coefficients and subject the extracted features to abundant of feature ranking and feature selection schemes to determine the best combination of features to maximize interclass similarity and assist in the union of classifiers, such as the support vector machine (SVM), sequential minimal optimization (SMO), random forest, and Bayes techniques.

.

II. METHOD

A. Extracting Data using OCR

The OCT macular scan report of a patient is taken as input in the form of .pdf or any other popular image format such as .jpeg, .jpg, .bmp etc.

This input is processed in a batch and extracts all the data present in it to a single text file. It uses pattern recognition and feature detection to recognize text in images. An OCR engine is used for this purpose. This paper uses one such engine to perform the stated task namely, TESSERACT. Tesseract was founded in 1985 by HP and is still said to be the most efficient and open source engine present in the world. It is a cross-platform engine but supports only English text efficiently. It needs to know about different shapes of the same character by having different fonts separated explicitly in the same folder where the input files are located. Tesseract is used to take the batch input of these high resolution OCT macular scans and generate the text data recognized by it in a file.

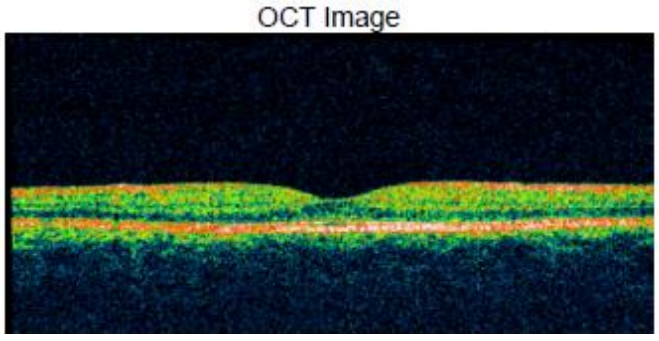

Fig.1. OCT Scan

The OCR technique based on Tesseract engine is used to extract specific information from the OCT Scans such as macular retinal thickness and its resolution, quality of scan etc. This is represented in Fig 2 and Fig 3.

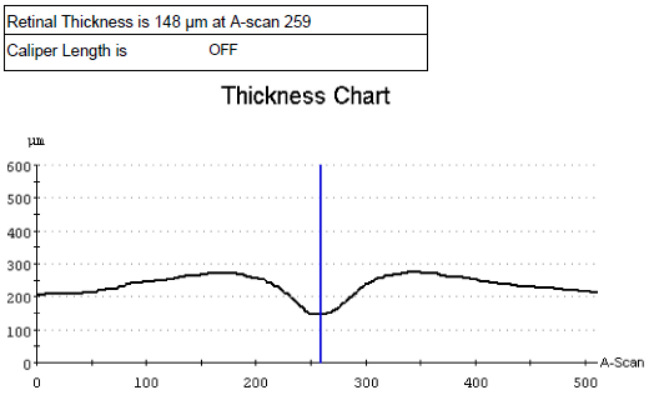

Fig.2. OCT Scan Measurements

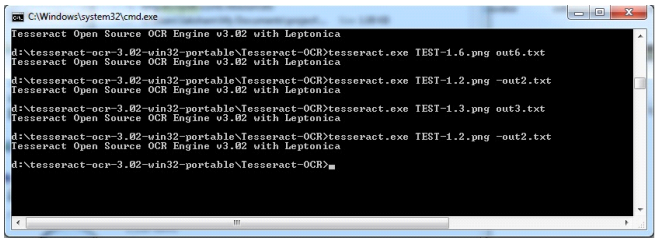

Fig.3. Use of Tesseract OCR Engine

B. Data Pre-Processing

Data preprocessing plays a very important part in many deep learning algorithms. In practice, many methods work best after the data has been normalized and whitened. Data normalization includes simple rescaling of data followed by mean subtraction and ultimately feature standardization. Many complex algorithms rely on whitening to learn good features. The most popular whitening is PCA/ZCA whitening which depends on the value of epsilon. So we must zero mean the features across the data set to ensure the product of the reciprocal of variance with each data equals zero. However, it is preferable to set epsilon to a value such that low-pass filtering is achieved. Java is used in this paper for pre processing the text file to extract the important information such as the coordinates of the macular thickness map and the resolution depth of it. Fig 4 explains the overall design of the process with the help of petri net diagram.

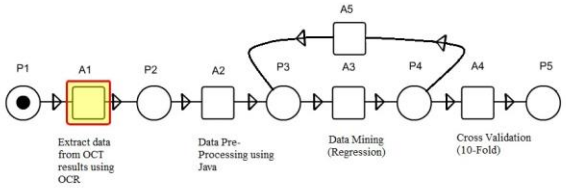

Fig.4. Process Overview

C. Data Mining

Data mining is a field of computer science which falls under the analysis step of the Knowledge Discovery in Databases process, or KDD. Apart from the raw analysis step, it includes database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Eventunating with data mining the first step consists of Selection of training data which is associated with the selection of wanted data using appropriate algorithm after that Data mining algorithm is applied.

The algorithm which has been used here is Regression algorithm which is defined as a model that predicts one or more continuous variables, such as profit or loss, based on other attributes in the dataset. The tool used for Regression algorithms is WEKA which is a powerful, yet easy to use tool for machine learning and data mining. The algorithms can either be applied directly to a dataset or called from one's own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association-rules, and visualization.

We can use various data mining tehniques such as classification, regression, clustering etc for the implementation part of this paper which are compared in [3]. It was found that the classification technique showed 73% efficiency tested over 20 samples [3] but we aim to use regression technique for this process and increase and contrast with the same.

The below XML MODEL shows the REGRESSION process-

<RegressionModel

functionName="regression"

modelName="Detection using linear regression"

modelType="linearRegression"

targetFieldName="VF values">

<RegressionTable intercept="172.34">

<NumericPredictor name="coordinate_x"

exponent="1" coefficient="5.10"/>

<NumericPredictor name="coordinate_y"

exponent="1" coefficient="0.01"/>

<NumericPredictor name="Depth_resolution"

exponent="1" coefficient="160"/>

<CategoricalPredictor name="defected"

value="-16, -16, -16, -16, -16, -2,

-20, -16, -25, -16, -10, -2,

-20, -25, -25, -16, -16,

-20, -20, -16, -16, -2,

-16, -16, -16, -16, -16, -2,

-16, -16, -16, -16, -16, -2"

coefficient="18.1"/>

</RegressionTable>

</RegressionModel>

The attribute's list contains the name of the variable, the value attribute, and the coefficient. This value of coefficient is then multiplied with the variables to get the resultant value. We use some methods like dropping irrelevant variables and dropping independent variables with small coefficient values to get better results and make the process more efficient. If the specified value of an independent value occurs in the table, the term variable_name is replaced with 1 and then the coefficient is multiplied by 1. If the value is missing in the table then the term variable_name is replaced with 0 and if the independent variable has missing values then the mean attribute is used to replace the missing values with the mean value.

The most common problem in data mining is the use of a regression equation for predicting the value of a dependent variable when one has a number of variables available to choose in a set as one or more independent variables in the model. So an efficient algorithm is needed for appropriate subset selection. These are the candidate algorithms which can be used in this process.

Forward Selection

In forward selection algorithm we keep adding variables one at a time to construct a reasonably good subset. The algorithm is as follows:

Algorithm Forward_selection

{

1. S = {constants};

repeat

2. if(var is not a subset of S)

compute reduction in sum of squares of residuals, R;

3. Let sigma square, $\hat{\sigma}^2$ be an unbiased estimate for the model consisting of the set S of variables. For some i, the largest reduction in SSR is-

$$F_i = Max_{i \notin S} \frac{SSR(S) - SSR(S \cup \{i\})}{\hat{\sigma}^2(S \cup \{i\})}$$

4. if(Fi > Fin) S = S U i ;

until i = n; }

Backward Elimination

In backward selection algorithm we keep eliminating variables one at a time to construct a reasonably good subset. It has the advantage of having all the variables in the set, S initially. The algorithm is as follows:

Algorithm Backward_elimination

{

1. S = {All var};

repeat

2. Let sigma square, $\hat{\sigma}^2$ be an unbiased estimate for the model consisting of the set S of variables. For some i, the smallest increase in SSR is-

$$F_i = Min_{i \notin S} \frac{SSR(S-\{i\}) - SSR(S)}{\hat{\sigma}^2(S)}$$

3. if (Fi < Fout)

S = S - i ;

until i = 0;

}

R^2 Method

Since R^2 is directly proportional to the sum of squares of residuals, so we pick the largest value of R^2. We can also choose the subset with maximum size (R^2_{adj}) which can be represented as -

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}(1-R^2).$$

We can also use other methods like step wise regression, all steps regression and [7] Mallow's C_p for subset selection. Mallow's C_p is a very popular and efficient method for subset selection. This value for C_p can be computed by the following formula:

$$C_p = \frac{SSR}{\hat{\sigma}^2_{Full}} + 2k - n,$$

These steps are compared and contrasted in a table.[4]

Fig 5 gives us the overview of the functions and provide us look and feel of Weka which we are using in this paper.

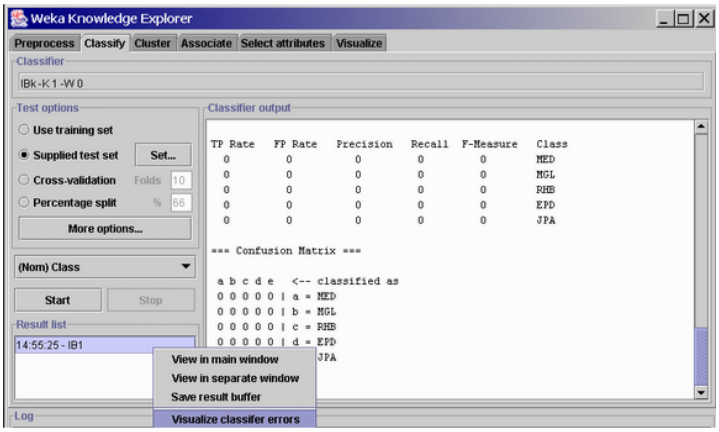

Fig.5. Weka Overview

D. Cross Validation

Cross-validation is a technique which is used to protect against the overfitting of data in a predictive model, mostly used in a case where the amount of data may be limited. In cross-validation, a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate. This paper uses k-fold cross validation technique. In this technique, a single sample of data is used as the validation data for testing model and the remaining k-1 for the training model.

The most popular k value used is 10 in which 10% of the sample data is used as test data and the rest 90% as the training data (Fig 6). This gives us a clear and better idea in analyzing the Statistics of the method proposed in the paper. The Cross-Validation error can be measured as-

$$CV(\lambda) = \frac{1}{K}\sum_{k=1}^{K} E_k(\lambda)$$

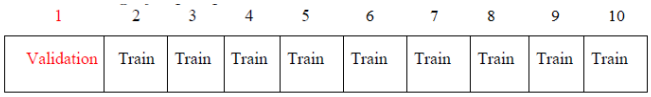

Fig.6. 10-fold Cross Validation

III. CONCLUSION

Glaucoma steals eye sight silently and causes damage to your eye's optic nerve and gets worse over time. There are no notice able symptoms of glaucoma and is incurable later. Little work has been done for detection of glaucoma in early

stages. However automated detection of glaucoma has proved to be successful method as compared to the other methods. This paper aims to use OCT images of 100 cases comprising both glaucomatic patients and healthy patients and implement data pre-processing with data mining and finally tests them using 10- fold cross validation to obtain a high success rate.

REFERENCES

[1] Khalil T and et all., "Review of Machine Learning Techniques for Glaucoma Detection and Prediction," Science and Information Conference of the IEEE , 2014.

[2] Saja Usman and et all., "A Review On Different Glaucoma Detection Methods, "International journal of advanced research in Engineering and technology, 2014.

[3] Kavita Choudhary and et all., "Glaucoma Detection using Cross Validation Algorithm," Fourth International Conference on Advanced Computing & Communication Technologies, pp. 478–482, 2014.

[4] Shabir et al., "Evaluation of Knowledge Extraction Using Various Classification Data Mining Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, pp. 251-256, 2013

[5] Juan Xu and et all.,"3D Optical Coherence Tomography Super Pixel with Machine Classifier Analysis for Glaucoma Detection", IEEE,2011.

[6] Wong A and et all., "Alignment of Confocal Scanning Laser Ophthalmoscopy Photoreceptor Images at Different Polarizations Using Complex Phase Relationships," Biomedical Engineering IEEE, 2009.

[7] Joshi A and et all., " Theoretic Feature Selection for Classification," American Control Conference ACC, 2007.

[8] Vermeer R.V and et all., "Modeling of scanning in Medical Imaging", IEEE,2006.

.